

ХИ-КВАДРАТ КРИТЕРИЙ ДЛЯ РЕГРЕССИОННОЙ МОДЕЛИ ПРИ НАЛИЧИИ КЛАССИФИКАЦИИ НАБЛЮДЕНИЙ

Е. С. Агеева, Ю. С. Харин

Белорусский государственный университет
Минск, Беларусь
e-mail: helenaageeva@gmail.com, kharin@bsu.by

Рассмотрены гипотезы согласия для регрессионной модели при наличии классификации наблюдений. Для проверки этих гипотез предложен специальный критерий на основе критерия хи-квадрат.

Ключевые слова: регрессия; классификация наблюдений; хи-квадрат критерий.

CHI-SQUARED TEST FOR REGRESSION MODEL UNDER CLASSIFICATION OF THE DEPENDENT VARIABLE

Н. Ageeva, Yu. Kharin

Belarusian State University
Minsk, Belarus

Goodness-of-fit hypotheses for regression model under classification of the dependent variable were considered. For hypothesis testing a modification of chi-squared test was suggested.

Keywords: regression; classification of observations; chi-squared test.

ВВЕДЕНИЕ

Регрессионной моделью описываются многие процессы, однако на практике зачастую данные наблюдаются с некоторыми отклонениями от классической модели [1]. Эта работа посвящена такому искажению, как группирование [2], а точнее одному из его частных случаев – классификации.

Важной задачей является правильный выбор функции регрессии. К сожалению, применять критерии для проверки гипотез согласия для классической регрессионной модели не удастся, так как при наличии классификации наблюдений отсутствуют остатки. Хи-квадрат критерий [3] не применим, поскольку наблюдения являются неравномерно распределенными. Однако нам удалось построить видоизмененный критерий хи-квадрат.

ПОСТАНОВКА ЗАДАЧИ

Пусть имеет место регрессионная модель:

$$Y_t = f^0(X_t) + \xi_t, \quad t = 1, \dots, n, \quad (1)$$

где n – объем выборки; $X_t \in X \subseteq R^N$ – наблюдаемый вектор-столбец регрессоров; $\xi_t \in R^1$ – ненаблюдаемая случайная величина ошибок с нормальным распределением вероятностей $N(0, (\sigma^0)^2)$ ($\{\xi_t\}_{t=1}^n$ – независимые в совокупности). Параметрами модели будут являться дисперсия ошибок $(\sigma^0)^2$ и неизвестная функция регрессии $f^0(\cdot)$.

Определена последовательность K непересекающихся интервалов ($2 \leq K < +\infty$):

$$A_k = (a_{k-1}, a_k], \quad k \in K = \{1, \dots, K\}, \quad a_0 = -\infty, \quad a_K = +\infty. \quad (2)$$

Эта система интервалов задает классификацию Y_t : Y_t относится к классу номер v_t , если $Y_t \in A_{v_t}$. Тогда вместо точных наблюдений $Y_1, \dots, Y_n$ наблюдаются лишь соответствующие номера классов $v_1, \dots, v_n \in K$.

Задача заключается в том, чтобы, зная набор интервалов $A_1, \dots, A_K$, значения регрессоров $\bar{X} = \{X_1, \dots, X_n\}$ и соответствующие номера классов $\bar{H} = \{v_1, \dots, v_n\}$, построить статистический тест для проверки гипотез о виде функции регрессии:

$$\begin{aligned} H_0: & f^0(\cdot) \in F, \\ H_1: & f^0(\cdot) \notin F, \end{aligned} \quad (3)$$

где $F = \{F(X; \theta), X \in X, \theta \in \Theta \subseteq R^m\}$ – предполагаемое параметрическое семейство функции регрессии. Например, требуется установить, можно ли использовать линейную регрессию для описания полученных данных. Заметим, что в случае, если верна гипотеза H_0 , существует такой параметр $\theta^0 \in \Theta$, что $f^0(\cdot) = F(\cdot; \theta^0)$. В этом случае параметром модели будет составной вектор $\delta = (\theta^T, \sigma^2)^T \in \Delta \subseteq R^{m+1}$, $\Delta \subseteq \Theta \times (0, +\infty)$.

ХИ-КВАДРАТ КРИТЕРИЙ ДЛЯ ПРОВЕРКИ ГИПОТЕЗ СОГЛАСИЯ

В случае верной гипотезы H_0 справедливо следующее утверждение:

$$P_{H_0}\{Y_t \in A_k\} = P_{H_0}\{v_t = k\} = P(k; \delta, X),$$

где $P(k; \delta, X) = \Phi\left(\frac{a_k - F(X; \theta)}{\sigma}\right) - \Phi\left(\frac{a_{k-1} - F(X; \theta)}{\sigma}\right)$, $k \in K$. Тогда истинный параметр $\delta^0 = ((\theta^0)^T, (\sigma^0)^2)^T \in \Delta$ можно оценить по методу максимального правдоподобия:

$$\hat{\delta} = \arg \max_{\delta \in \Delta} \sum_{t=1}^n \ln P(v_t; \delta, X_t). \quad (4)$$

В [4] приведены условия сильной состоятельности и асимптотической нормальности для ОМП (4).

Введем следующие обозначения:

$$fr(k) = \sum_{t=1}^n I(v_t = k), \quad p_k(\delta) = \frac{1}{n} \sum_{t=1}^n P(v_t; \delta, X_t),$$

$$D(\delta) = (d_{ij}(\delta))_{i,j=1}^{K-1}, \quad d_{ij}(\delta) = I(i=j) \cdot p_k(\delta) - \frac{1}{n} \sum_{t=1}^n P(i; \delta, X_t) P(j; \delta, X_t),$$

где $I(A)$ – индикаторная функция. Рассмотрим случайные величины $\sqrt{n}(\frac{1}{n} \bar{f}_r(k) - p_k(\delta))$, $k = 1, \dots, K-1$. Заметим, что при верной гипотезе H_0 выполнено

$$E_{H_0} \left\{ \left(\sqrt{n} \left(\frac{1}{n} \bar{f}_r(k) - p_k(\delta^0) \right) \right)_{k=1}^{K-1} \right\} = 0_{K-1}, \quad \text{Var}_{H_0} \left\{ \left(\sqrt{n} \left(\frac{1}{n} \bar{f}_r(k) - p_k(\delta) \right) \right)_{k=1}^{K-1} \right\} = D(\delta^0).$$

Введем следующее обозначение:

$$\bar{\chi}^2(\delta, \bar{X}, \bar{H}) = \left(\left(\sqrt{n} \left(\frac{1}{n} \bar{f}_r(k) - p_k(\delta) \right) \right)_{k=1}^{K-1} \right)^T (D(\delta))^{-1} \left(\sqrt{n} \left(\frac{1}{n} \bar{f}_r(k) - p_k(\delta) \right) \right)_{k=1}^{K-1}.$$

Рассмотрим информационную матрицу Фишера:

$$\bar{\Gamma}_n(\delta) = \frac{1}{n} \sum_{t=1}^n E_{\delta} \{ \nabla_{\delta} \ln P(v_t, \delta, X_t) (\nabla_{\delta} \ln P(v_t, \delta, X_t))^T \},$$

где ∇f – это градиент функции f , и матрицу

$$P^1(\delta) = \left(\frac{\partial p_k(\delta)}{\partial \delta_i} \right)_{k=1, \dots, K-1; i=1, \dots, m+1}.$$

Теорема 1. Пусть функция $F(X; \theta^0)$ ограничена на X . Тогда

$$L_{H_0} \{ \bar{\chi}^2(\delta^0, \bar{X}, \bar{H}) \} \xrightarrow{n \rightarrow \infty} \chi_{K-1}^2.$$

Теорема 2. Пусть выполнены следующие условия:

1. ОМП $\hat{\delta}$ является состоятельной оценкой вектора параметров δ^0 .
 2. Для любого $\theta \in \Theta$ функции $F(X; \theta)$, $\frac{\partial F(X; \theta)}{\partial \theta_i}$, $\frac{\partial^2 F(X; \theta)}{\partial \theta_i \partial \theta_j}$, $\frac{\partial^2 F(X; \theta)}{\partial \theta_i \partial \theta_j \partial \theta_s}$, $i, j, s = 1, \dots, m$, ограничены на X .
 3. Матрица $\bar{\Gamma}_n(\delta)$ положительно определена: $\bar{\Gamma}_n(\delta^0) > 0$.
 4. Для матрицы $\bar{\Gamma}_n(\delta^0)$ выполнено $\lim_{n \rightarrow \infty} |\bar{\Gamma}_n(\delta^0)| = d > 0$.
 5. Первые $K-3$ собственные значения матрицы $I_{K-1} - (D(\delta^0))^{-1} P^1(\delta^0) (\bar{\Gamma}_n(\delta^0))^{-1} (P^1(\delta^0))^T$, где I_{K-1} – единичная матрица размерности $K-1$, стремятся к единице, а остальные к нулю.
- Тогда

$$L_{H_0} \{ \bar{\chi}^2(\hat{\delta}, \bar{X}, \bar{H}) \} \xrightarrow{n \rightarrow \infty} \chi_{K-3}^2.$$

Замечание. Число степеней свободы в теореме 2 обусловлено тем, что случайная величина v_t зависит от двух параметров: математического ожидания $\mu_t = f^0(X_t)$ и дисперсии $(\sigma^0)^2$.

КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

Рассмотрим случай одномерных регрессоров. В качестве предполагаемого параметрического семейства функции регрессии возьмем множество линейных функций:

$$F = \{ \theta_1 + \theta_2 X, X \in X, \theta = (\theta_1, \theta_2)^T \in \Theta \subseteq R^2 \}.$$

Для проверки гипотез (3) используем следующее решающее правило:

$$d(\bar{X}, \bar{H}) = \begin{cases} 0, & \text{если } \bar{\chi}^2(\hat{\delta}) \leq F_{\chi_{K-3}^2}^{-1}(1-\varepsilon), \\ 1, & \text{иначе,} \end{cases} \quad (5)$$

где ε – заданный уровень значимости, $F_{\chi_{K-3}^2}^{-1}(1-\varepsilon)$ – квантиль уровня $1-\varepsilon$ распределения χ_{K-3}^2 .

Предполагаем, что $K = 5$, $A_1 = (-\infty, 30)$, $A_2 = [30, 36)$, $A_3 = (36, 44]$, $A_4 = (44, 50]$, $A_5 = (50, +\infty)$, $(\sigma^0)^2 = 4$. Регрессоры равномерно распределены на $[0, 10]$. Для каждого объема выборки n проводим две серии по $Q=1000$ экспериментов: в первой серии в качестве истинной функции регрессии используем функцию $f_1(x) = 20 + 4x$ и генерируем данные $\bar{X}_{1,q}, \bar{H}_{1,q}$, $q = 1, \dots, Q$; во второй серии в качестве истинной функции регрессии используем функцию $f_2(x) = 27,5 + 0,4x^2$ и генерируем данные $\bar{X}_{2,q}, \bar{H}_{2,q}$, $q = 1, \dots, Q$.

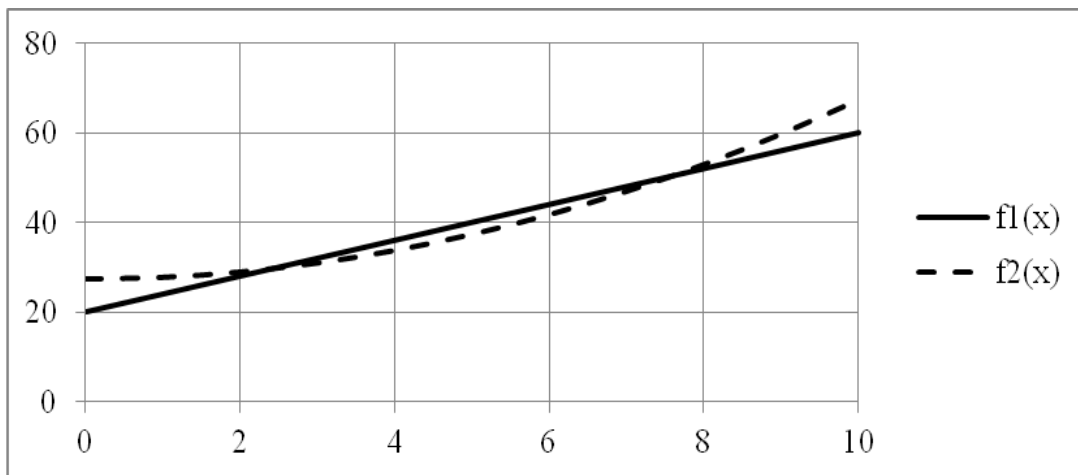


Рис. 1. Графики функций $f_1(x)$ и $f_2(x)$

Предполагаем уровень значимости $\varepsilon = 0,05$. Вычислим статистики

$$\alpha = \frac{1}{Q} \sum_{q=1}^Q d(\bar{X}_{2,q}, \bar{H}_{2,q}), \quad w = \frac{1}{Q} \sum_{q=1}^Q d(\bar{X}_{1,q}, \bar{H}_{1,q}).$$

Статистика α является оценкой вероятности ошибки I рода для теста (5). Статистика w является оценкой мощности для теста (5). На рис. 2 и 3 представлены графики зависимостей α и w от объема выборки n соответственно.

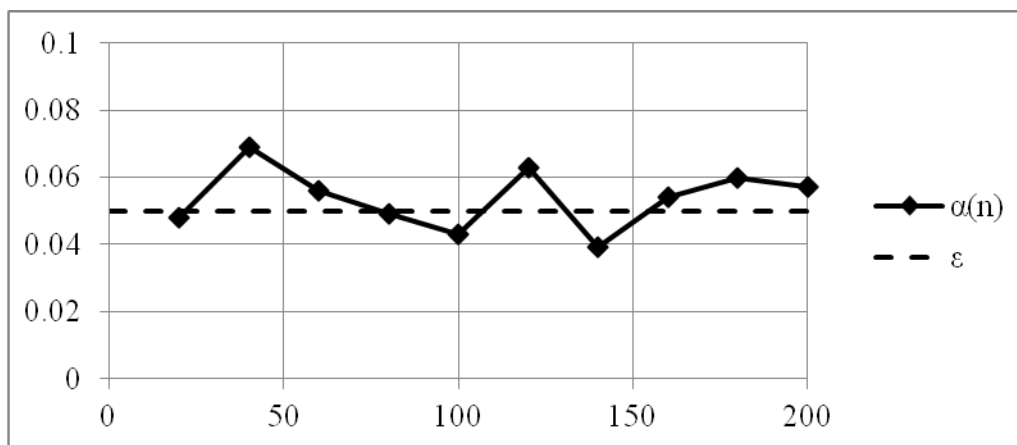


Рис. 2. Графики зависимостей α от n

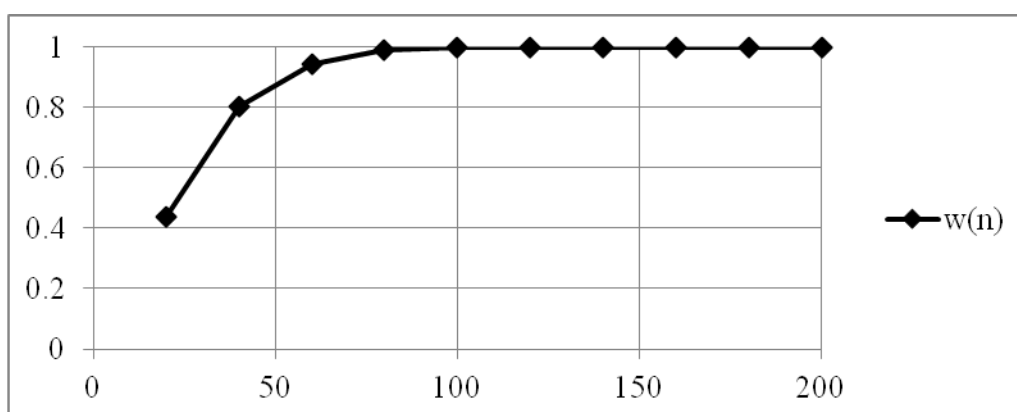


Рис. 3. Графики зависимостей w от n

ЗАКЛЮЧЕНИЕ

Из численных экспериментов видно, что вероятность ошибки первого рода теста (5) колеблется около заданного уровня значимости ε , его мощность при этом быстро стремится к 1.

Также следует отметить, что предложенный критерий можно применять для проверки гипотез согласия для классической регрессионной модели.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Kharin Yu. Robustness in statistical forecasting. Springer, 2013.
2. Heitjan D. F. Inference from Grouped Continuous Data: A Review // Statistical Science. 1989. Vol. 4. № 2. P. 164–183.
3. Крамер Г. Математические методы статистики. М. : Мир, 1975.
4. Ageeva H., Kharin Yu. ML estimation of multiple regression parameters under classification of the dependent variable // Lithuanian Mathematical J. 2015. Vol. 55. № 1. P. 48–60.